AD_____

GRANT NUMBER DAMD17-94-J-4332

TITLE: Statistical Methods for Analyzing Time-Dependent Events in Breast Cancer Chemoprevention Studies

PRINCIPAL INVESTIGATOR: George Y. C. Wong, Ph.D.

CONTRACTING ORGANIZATION: Strang Cancer Prevention Center
New York, New York 10021

REPORT DATE: November 1997

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for public release;
distribution unlimited

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	November 1997	Annual (30 Sep 96 - 29 Sep 97)

4. TITLE AND SUBTITLE

Statistical Methods for Analyzing Time-Dependent Events in Breast Cancer Chemoprevention Studies

5. FUNDING NUMBERS

DAMD17-94-J-4332

6. AUTHOR(S)

George Y.C. Wong, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Strang Cancer Prevention Center
New York, New York 10021

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The overall aim of our research proposal is the statistical inference of nonparametric estimates, the redistribution-to-the-inside estimator (RTIE) and the generalized maximum likelihood estimator (GMLE), for the survival function of a time-to-event variable that is subject to interval censoring. The RTIE, which is proposed by us, has a closed-form expression and is equal to the GMLE under a homogeneous condition. The GMLE is the standard optimal procedure in survival analysis. However, no closed-form expression for the GMLE is available, and asymptotic distribution theory for it has been limited. Our research efforts in the third year have focused on the asymptotic inference of the GMLE under conditions more general than the discrete distribution assumption that we previously imposed on the censoring variables. Additionally, we have derived an asymptotic nonparametric two-sample test procedure for comparing two populations. Finally, we have begun investigating the asymptotic inference of Cox regression model for interval-censored data by establishing consistency of the GMLE of the model parameters under finite assumptions on both the survival and censoring distributions.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Breast Cancer , Interval Censorship, Consistency, Asymptotic Normality	12
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

____ Where copyrighted material is quoted, permission has been obtained to use such material.

____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

____ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and Use of Laboratory Animals of the Institute of Laboratory Resources, National Research Council (NIH Publication No. 86-23, Revised 1985).

____ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

____ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

____ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

Nov 25, 1997

PI - Signature Date

A. TABLE OF CONTENTS

B. INTRODUCTION

In clinical follow-up studies, subjects are monitored at regular time intervals for a physical condition. It is often the case that an event under observation can take place in between two successive visits, and it may not be possible for the subject to know the time to such an event exactly. For example, consider the situation in which a group of women at high risk for breast cancer is asked to take a chemopreventive substance for a fixed time period. At the end of the period, each participating woman is required to submit a blood or urine sample at regular intervals in order to monitor the level of a validated intermediate biomarker. Let X denote the time from cessation of use of the agent to the loss of its protective effect, quantified as a return to baseline value of the biomarker. If a woman submits a sample for assay on a daily basis, the value of X can be observed exactly, unless the protective effect is still present by the time the study is terminated so that X is right censored in the usual sense of survival analysis. In practice, however, the follow-up interval can be a week or longer; therefore the exact value of X is generally unknown but is known to lie between the time points L and R, where L is the number of days from cessation of agent intake to the last time the sample was assayed and the protective effect was still present, and R is the number of days from cessation of agent intake to the most recent time the sample was assayed. If the protective effect is still present, then R takes the value infinity. In any case, when the value of X is only known to lie between (L, R), we say that X is censored in the interval (L, R). Therefore the observed data consist of either censoring intervals (L, R) or exact observations $X = L = R$.

We consider nonparametric estimation of the distribution function $F(t)$ of a real-valued random variable X (or its survival function $S(t) = 1 - F(t)$, where $F(t)$ =P$\{X \leq t\}$), when the sample data are incomplete due to restricted observation brought about by interval censoring.

At present, there are only two estimation procedures of S for interval-censored (IC) data that are generalized maximum likelihood estimates (GMLE) in the sense of Kiefer and Wolfowitz [1]. The first one is due to Peto [2] and makes use of the Newton-Ralphon algorithm. The second is due to Turnbull [3] and makes use of a self-consistent algorithm. A solution to the latter algorithm is called a self-consistent estimator (SCE) of S. In each case, there is no closed-form expression for the estimator.

In the first year of our research, we focused our attention on IC data that satisfy a condition which we called DI condition: IC data $\{L_1, R_1\}, ..., \{L_n, R_n\}$ are said to satisfy DI condition if given any two censoring intervals, either they are disjoint or one is a subset of the other. In a clinical study in which every subject has the same follow-up schedule, say at time point a_1, a_2, ..., a_k, then $\{L, R\} = \{0, a_1\}$, or $\{a_i, a_{i+1}\}$ or $\{a_i, \infty\}$, and hence such interval-censoring data will satisfy DI Condition.

Under DI interval-censorship model, we extended Efron's [4] redistribution -to-the-right idea for right-censored data and proposed a redistribution-to-the-inside method to yield a nonparametric estimator of $S(t)$ which we called redistribution-to-the-inside estimator (RTIE). Such an estimate has a closed-form expression and can be readily calculated for IC data of any size. The availability of an explicit expression for the RTIE has enabled us to show that it is the GMLE under DI condition, and to establish asymptotic properties of the RTIE.

More often than not, IC data do not satisfy the DI condition. In a clinical follow-up situation, for example, a patient may miss a scheduled appointment. Therefore, it is necessary to consider asymptotic inference under more general conditions of interval censorship.

There are 4 situations in which IC data can occur.

Case 2 IC data (C2 data) consist of right-, left- and strictly interval-censored observations. Here, an observation is called *right censored* if $R = \infty$, *left censored* if $L = 0$, *exact* if $L = R$ and *strictly interval censored* if $0 < L < R < \infty$. Examples of C2 data can be found in [5]. Mixed IC data (MIC data) consist of both C2 data and exact observations, referred to as partially IC data in our second-year report. Yu, Li and Wong [6] presented a set of MIC data in breast cancer research. Doubly-censored data (DC data) consist of right-, left-censored and exact observations. Examples of DC data can be found in [7]. Case 1 IC data (C1 data)) consist of right-censored and left-censored observations. Examples of C1 data can be found in [8] and [9].

Four different interval censorship models corresponding to the four different types of IC data have been proposed. They are the C2 model, the mixture interval censorship model (MIC model), the DC model and the C1 model.

To study the asymptotic properties of the GMLE, we make use of the following assumptions:

(AS1) The censoring distribution is discrete but the survival distribution is arbitrary.

(AS2) The support set of the censoring vector is finite, but the survival distribution is arbitrary.

(AS3) A probability restriction. See Section C.

(AS4) A probability restriction. See Section C.

(AS5) The censoring distribution and the survival distribution are arbitrary, but have to satisfy some regularity conditions.

In our second year of research, we established the following important asymptotic results of the GMLE under both DI and non-DI conditions:

1. Under the C1 model or the C2 model, the GMLE is strongly consistent under assumption (AS1)
2. Under the C1 model, the GMLE is asymptotically normal and efficient under Assumption (AS1).
3. Under the C2 model, the GMLE is asymptotically normal and efficient under assumption (AS2)
4. Under the MIC model, both the SCE and the GMLE are strongly consistent under Assumption (AS2).
5. Under the MIC model, both the SCE and the GMLE are asymptotically normal and efficient under Assumption (AS1).

In our third year, we established the following important asymptotic results for the GMLE under both DI and non-DI conditions:

6. Under the MIC model, the SCE and the GMLE are strongly consistent under Assumptions (AS3).
7. Under the MIC model, the SCE and the GMLE are asymptotically normal and efficient under Assumptions (AS3) and (AS4).

8. Under the DC model, we have proposed a modified GMLE and proved that the estimator is strongly consistent, asymptotically normal and efficient under Assumption (AS5).
9. We have constructed an asymptotic nonparametric two-sample test procedure for all types of IC data, and applied it to a breast cancer relapse follow-up study.
10. We have proved the consistency of the GMLE of the parameters in the proportional hazards model for C2 and MIC data.

C. BODY

C.1. Case 1 and Case 2 Models.

We established consistency and asymptotic normality of the GMLE under assumptions (AS1) or (AS2) in our second year of research.

In the third year, we have revised two manuscripts pertaining to these results and they are now accepted by two peer-reviewed statistical journals ([10] and [11]).

C.2. MIC Model.

In the second year of our research, we proposed for MIC data the MIC model, which is a mixture of the C2 interval censorship model and the usual right censorship (RC) model (see Yu, Li and Wong [6], [12] and [13]). The C2 model assumes that X and the random censoring vector (Y, Z) are independent and that $Y < Z$ with probability one. The RC model assumes that there is a random censoring time T, which is independent of X, and the observable information from the RC model is $(\min(X,T), I(X \leq T))$. We introduce a random variable D to distinguish failure times coming from the two models:

$$D = \begin{cases} 1 & \text{if the observation is from the RC model} \\ 0 & \text{if the observation is from the C2 model.} \end{cases}$$

Let $P\{D = 1\} = \pi$, where $0 < \pi \leq 1$. Formally, a MIC data point is regarded as an observation from the RC model with probability π and from the C2 model with probability $1 - \pi$.

To express observed MIC data as intervals, we introduce a notation $\lfloor L, R \rfloor$ defined as follows:

$$\lfloor L, R \rfloor = \begin{cases} [0, Y) & \text{if } D = 0 \text{ and } X < Y \\ [Y, Z) & \text{if } D = 0 \text{ and } Y \leq X < Z \\ [Z, \infty) & \text{if } D = 0 \text{ and } X \geq Z \\ (T, \infty) & \text{if } D = 1 \text{ and } X > T \\ [X, X] & \text{if } D = 1 \text{ and } X \leq T, \end{cases}$$

where $[X, X]$ is an exact observation. Let (L_i, R_i), $i = 1, 2, \ldots, n$ be a random sample from the random vector (L, R), and $\lfloor l, r \rfloor$ be a realization of $\lfloor L, R \rfloor$. We say that the MIC data $\lfloor L, R \rfloor$ are from a *mixture* interval censorship model (MIC model).

Define $\tau = \sup\{t;\ P\{\min(X,T) \leq t\} < 1\}$, $\tau_Y = \sup\{t;\ P\{Y \leq t\} = 0\}$. and $\tau_Z = \sup\{t;\ P\{Z \leq t\} < 1\}$. We assume that $\tau \geq \tau_Z$.

(AS3) $\mathrm{P}\{L = \tau\} > 0$ if $P(X < \tau) < 1$ and $\mathrm{P}\{R = \tau_Y\} > 0$ if $P(X \leq \tau_Y) > 0$.

Theorem 1. *Under assumptions (AS2) and (AS3), the SCE $\hat{F}(x)$ satisfies that*

$\lim_{n\to\infty} \sup_{x\le\tau} |\hat{F}(x) - F(x)| = 0$ *a.s.*

To establish asymptotic normality for the SCE, we need an additional assumption on the distribution function, namely,

(AS4) $P\{X \in I_i \cap I_j\} > 0$ for any two realizations, $\lfloor L_i, R_i \rfloor = I_i$ and $\lfloor L_j, R_j \rfloor = I_j$, of $\lfloor L, R \rfloor$, provided $I_i \cap I_j \neq \emptyset$.

Theorem 2. *Under Assumptions (AS2), (AS3) and (AS4), the SCE $\hat{F}(x)$ satisfies that for $x \le \tau$, $\sqrt{n}(\hat{F}(x) - F(x)) \xrightarrow{\mathcal{D}}$ a normal random variable as $n \to \infty$.*

The above two theorems are summarized in two separate papers. In the third year of our research, we have revised these two papers and they are now accepted by two peer-reviewed journals ([6] and [13]). Moreover, in the third year of our research, we relaxed the assumptions in these two theorems and prove the following results.

Theorem 3. *Under assumption (AS3), the SCE $\hat{F}(x)$ satisfies that*

$\lim_{n\to\infty} \sup_{x\le\tau} |\hat{F}(x) - F(x)| = 0$ *a.s.*

Theorem 4. *Under Assumptions (AS3) and (AS4), the SCE $\hat{F}(x)$ satisfies that for $x \le \tau$, $\sqrt{n}(\hat{F}(x) - F(x)) \xrightarrow{\mathcal{D}}$ a normal random variable as $n \to \infty$.*

These results are extensions of the results we obtained in [6] and [13] by deleting assumption (AS2). A manuscript [12] pertaining to these result has been submitted for publication.

C.3. DC Model.

We consider efficient estimation of a survival function S of the random variable X with doubly-censored data. The double censorship model assumes that X and the random vector (Y, Z) are independent and $Y < Z$ with probability one, and that X is uncensored if $Y < X \le Z$, right censored if $Z < X$ and left censored if $X \le Y$. Let S_Z and S_Y be the survival functions of Z and Y, respectively, and let $K = S_Y - S_Z$. Under the assumption

(AS5) $\qquad K(x-) > 0$ for all x such that $S(x) < 1$ and $S(x-) > 0$,

we present an example in [14] to demonstrate that the GMLE of S is not asymptotically normally distributed and is not asymptotically efficient, and we propose a modified GMLE (for details, see [14]) and establish the following results for such a modified estimator.

Theorem 5. *Under the DC model and Assumptions (AS5) and (AS3), the modified GMLE $\hat{F}(x)$ converges to $F(x)$ a.s. for all $x = a_i$, $i \ge 1$.*

Theorem 6. *Under the DC model and Assumptions (AS5), (AS3) and (AS4), the modified GMLE $\hat{F}$ satisfies $\sqrt{n}[\hat{F}(x) - F(x)] \xrightarrow{\mathcal{D}}$ a random normal variable as $n \to \infty$ for $x = a_i$.*

These two theorems are summarized in a paper [14] submitted for publication.

C.4. Two-Sample Nonparametric Test.

Based on the asymptotic results that we have established so far, we propose a test statistic that is represented as a cumulative weighted difference in the GMLE of distribution functions to test for their equality. The test statistic is applicable to all two types of IC data. We apply it to a breast cancer relapse follow-up study described as follows.

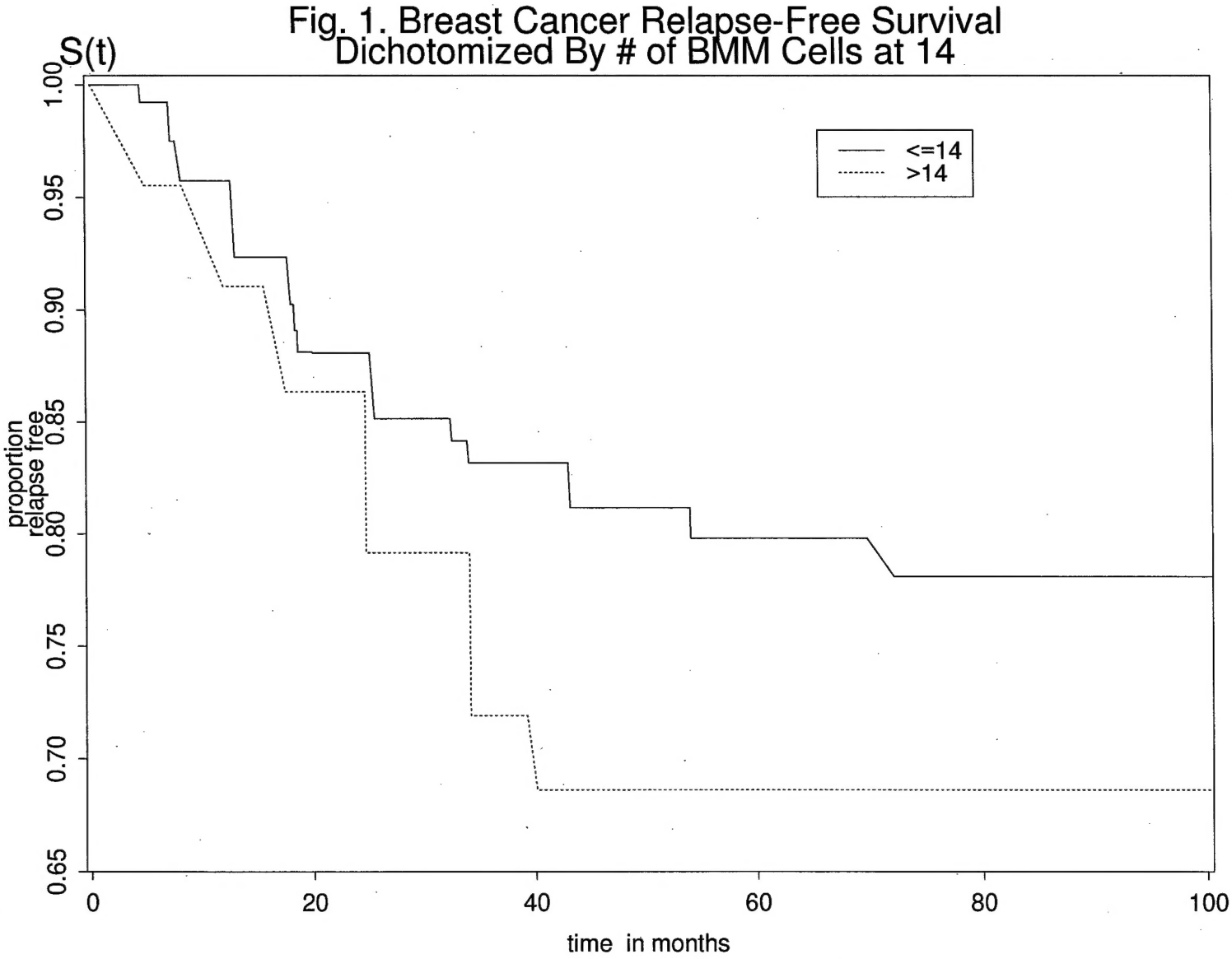
Fig. 1. Breast Cancer Relapse-Free Survival
Dichotomized By # of BMM Cells at 14
S(t)
proportion relapse free
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
<=14
>14
0
20
40
60
80
100
time in months

An Example Three hundred and seventy-four women with stages I - III unilateral invasive breast cancer surgically treated at Memorial Sloan-Kettering Cancer Center between 1985 and 1990 were followed for relapse. The median follow-up duration was 46 months. Relapse time was given by the time interval between surgery and the initial relapse. A relapse that took place between two successive follow-up visits was regarded as interval censored. If a patient did not relapse towards the end of the study, then her relapse time was right censored. Of the 374 observations, 300 were right censored (no relapse), 21 were left censored and 53 were strictly interval censored (74 relapses).

Bone marrow micrometastasis (BMM) was determined for each woman at the time of surgery. An important clinic question is whether remission duration is related to the number of BMM cells detected. Figure 1 compares the relapse-free GMLE curves of patients with number of BMM ≤ 14 versus those with number of BMM > 14. Our two-sample asymptotic nonparametric test yielded a P value close to 0.1. A manuscript [15] pertaining to the proposed asymptotic two-sample test procedure has been submitted to a statistical journal.

C.5. Proportional Hazards Model.

Under the restrictive assumptions that both X and the censoring vector take on finitely many values, we have proved that the GMLE of the parameters in Cox regression are consistent. However, we have not yet established asymptotic normality for the distributions of these parameter estimators. We have completed a first draft on the consistency result in [16].

D. CONCLUSIONS

In the third year of our DOD grant, we have essentially completed our research on the asymptotic inference on the GMLE of the survival function for IC data, including consistency, asymptotic normality and asymptotic efficiency. The results which we have established provide a set of fundamentally important statistical tools for the analysis of most types of IC data that are encountered in clinical follow-up studies. In the fourth and final year of our DOD grant, we plan to pursue the following three concluding issues on univariate IC data:

1. establish asymptotic results without the discrete assumption on the distribution of the random censoring vector,
2. construct counterparts of log-rank tests for IC data and derive their asymptotic properties,
3. establish asymptotic results for Cox regression parameters under conditions that are more relaxed than the requirement of finiteness.

We should point our that although we are administratively on the DOD grant in our fourth and final year, we are not receiving any funding from the DOD in this year. We will make every effort to complete our proposed work to bring our DOD funded research on the asymptotic survival analysis of univariate interval-censored data to a definitively satisfactory closure.

E. REFERENCES

[1] Kiefer, J and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* 27, 887-906.

[2] Peto, R. (1973). Experimental survival curves for interval-censored data. *Appl. Statist.* 22, 86-91.

[3] Turnbull, B. W. (1976). The empirical distribution function with arbitrary grouped, censored and truncated data. *J. Roy. Statist. Soc. Ser.* B, 38, 290-295.

[4] Efron, B (1967). The two sample problem with censored data. *Fifth Berkeley Symposium on Mathematical Statistics.* University of California Press, 831-853.

[5] Finkelstein, D.M. and Wolfe, R.A. (1985). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics*, 41, 933-945.

[6] Yu, Q., Li, L. and Wong, G. Y. C. (1997). On consistency of the self-consistent estimator of survival functions with interval censored data. *Scan. J. of Statist.* (Tentatively accepted).

[7] Leiderman, P.H., Babu, D., Kagia, J., Kraemer, H.C. and Leiderman, G.F. (1973). African infant precocity and some social influences during the first year. *Nature*, 242, 247-249.

[8] Ayer, M., Brunk, H.D., Ewing, G.M., Reid, W.T. and Silverman, E. (1955). An empirical distribution function for sampling incomplete information. *Ann. Math. Statist.* 26, 641-647.

[9] Keiding. N. (1991) Age-specific incidence and prevalence: A statistical perspective (with discussion) *JRSS*, A. 154, 371-412.

[10] Yu, Q., Schick, A., Li, L. and Wong, G. (1997). Asymptotic properties of the GMLE in the case 1 interval-censorship model with discrete inspection times. *Canadian Journal of Statistics.* (In press).

[11] Yu, Q., Schick, A., Li, L. and Wong, G. Y. C. (1997). Asymptotic properties of the GMLE of a survival function with case 2 interval-censored data. *Prob. & Statist. Letters.* (In press).

[12] Yu, Q., Li, L. and Wong, G. Y. C. (1997). Asymptotic properties of NPMLE with mixed interval-censored data. (Submitted to Annals of Statistics.)

[13] Yu, Q., Li, L. and Wong, G. Y. C. (1997). Asymptotic variance of the GMLE of a survival function with interval-censored data. *Sankhya,* A. (Accepted).

[14] Yu, Q., Li, L. and Wong, G. (1997). A modified GMLE with doubly-censored data. (Submitted to *Scan. J. of Statist.*).

[15] Wong, G. Y. C. and Q. Yu (1997). A generalized two-sample distance test with interval-censored data. (Submitted to *Statistics in Medicine*).

[16] Li, L., Yu, Q. and Wong, G. Y. C. (1997). Proportional hazard model with interval-censored and exact observations. (Under preparation).

F. APPENDICES

1. Manuscripts accepted by peer-reviewed journals (5):

[a] Yu, Q., Li, L. and Wong, G. Y. C. (1997). On consistency of the self-consistent estimator of survival functions with interval censored data. *Scan. J. of Statist.* (Tentatively accepted).

[b] Yu, Q., Schick, A., Li, L. and Wong, G. Y. C. (1997). Asymptotic properties of the GMLE in the case 1 interval-censorship model with discrete inspection times. *Canadian Journal of Statistics.* (In press).

[c] Yu, Q., Li, L. and Wong, G. Y. C. (1997). Asymptotic variance of the GMLE of a survival function with interval-censored data. *Sankhya,* A. (In press).

[d] Yu, Q., Schick, A., Li, L. and Wong, G. Y. C. (1997). Asymptotic properties of the GMLE of a survival function with case 2 interval-censored data. *Prob. & Statist. Let.* (In press).

[e] Yu, Q. and Wong, G. Y. C. (1997). Consistency of self-consistent estimators of a discrete distribution function with bivariate right-censored data. *Commun. Statist.* (In press).

2. Manuscripts submitted (3):

[f] Yu, Q., Li, L. and Wong, G. Y. C. (1997). Asymptotic properties of NPMLE with mixed interval-censored data. (Submitted to Annals of Statistics.)

[g] Yu, Q. and Wong, G. Y. C. (1997). A modified GMLE with doubly-censored data. (Submitted to *Scan. J. of Statist.*).

[h] Wong, G. Y. C. and Yu, Q. (1997). A generalized two-sample distance test with interval-censored data. (Submitted to *Statistics in Medicine*).

3. Manuscript under preparation (1):

[i] Li, L., Yu, Q. and Wong, G. Y. C. (1997). Proportional hazard model with interval-censored and exact observations. (Under preparation).

4. A list of conferences attended

ICSA 1997 Applied Statistics Symposium

May 30 - June 1, 1997

Rutgers University, New Jersey, USA

Title: Asymptotic Properties Of Self-Consistent Estimators of A Survival Function

by Qiqing Yu and George Y. C. Wong.